Psychological Bulletin
1988, Vol. 104, No. 3, 396-404


0033-2909/88/$00.75

# Heterogeneity of Variance in Experimental Studies: A Challenge to Conventional Interpretations

Anthony S. Bryk
Department of Education, University of Chicago

Stephen W. Raudenbush
Graduate School of Education, Michigan State University

The presence of heterogeneity of variance across groups indicates that the standard statistical model for treatment effects no longer applies. Specifically, the assumption that treatments add a constant to each subject's development fails. An alternative model is required to represent how treatment effects are distributed across individuals. We develop in this article a simple statistical model to demonstrate the link between heterogeneity of variance and random treatment effects. Next, we illustrate with results from two previously published studies how a failure to recognize the substantive importance of heterogeneity of variance obscured significant results present in these data. The article concludes with a review and synthesis of techniques for modeling variances. Although these methods have been well established in the statistical literature, they are not widely known by social and behavioral scientists.

Psychological researchers have tended historically to view heterogeneity of variance as a methodological nuisance, an unwelcome obstacle in the pursuit of inferences about the effects of treatments on means. In their discussion of variance heterogeneity, standard texts concentrate on identifying conditions under which such heterogeneity can safely be ignored so that standard analyses of means may proceed. It is usually argued that heterogeneity can be ignored when statistical tests for means are robust to violation of the homogeneity assumption (Glass & Hopkins, 1984, pp. 238–240; Hays, 1981, p. 287; Winer, 1971, pp. 37–39). When such violations cannot be ignored, analysts tend to assume heterogeneity must be eliminated. The primary strategy for eliminating heterogeneity is to find a transformation of the dependent variable that stabilizes treatment group variances, enabling retention of the homogeneity hypothesis (Kirk, 1982, pp. 79–84; Winer, 1971, pp. 397–402).

There has been little discussion in the literature of the causes of heterogeneity in experimental studies. Light and Smith (1971) noted that heterogeneity is likely to occur in program evaluation studies. Although providing good exploratory data analysis advice for examining heterogeneity of variance across groups, their focus remained fastened on making appropriate inferences about mean differences between programs.

In our view, to restrict considerations of variance heterogeneity to its effect on inferences about means is fundamentally misguided. We show in this article that the presence of heterogeneity of variance across groups in experimental studies indicates that treatments have differential effects across individuals. Rather than being a nuisance factor to be adjusted away, the presence of heterogeneity of variance is important empirical evidence of an interaction of treatments with some unspecified subject characteristics. To ignore variance heterogeneity, then, is tantamount to interpreting main effects while concealing significant interaction effects. Although it is generally understood that inferences about main effects are often misleading in the presence of interaction effects, ironically, we commit exactly the same error when we ignore heterogeneity of variance in experimental studies.

Further, in many cases, the nature of these differential effects is substantively interesting and can be crucial to evaluating the efficacy of treatments (e.g., see Bloom, 1984; Bryk, 1978). A common example of this phenomenon occurs when an experimental treatment has an effect on some subjects but not on others. This can result from technical problems in applying treatments, or it can result from differential responsiveness of subjects to the treatment. In either case, the treatment both increases the variance and affects the mean.

We present in this article a simple mathematical model that demonstrates how individual differences in treatment effects produce heterogeneity of variance across groups. On the basis of this model, we illustrate, with data from two previously published studies, a general strategy for examining the results of experiments when heterogeneity is present. The article concludes with a brief review of the necessary statistical theory for estimating treatment effects on variance and for testing hypotheses about these variance effects.

## Model of Treatment Effects

The simplest experimental design consists of random assignment of individuals to either a treatment group or a control group. The traditional statistical model corresponding to this design is

We wish to acknowledge support for this project from the Spencer Foundation and the Benton Center for Curriculum and Instruction at the University of Chicago. We also wish to thank the reviewers and Lincoln Moses for helpful comments on an earlier version of this manuscript.

Correspondence concerning this article should be addressed to Anthony S. Bryk, Department of Education, University of Chicago, 5835 South Kimbark Avenue, Chicago, Illinois 60637.

$$Y_i^C = \mu + e_i^C,$$
$$Y_i^E = \mu + \alpha + e_i^E. \quad (1)$$

Here, $Y_i^C$ and $Y_i^E$ represent outcome scores for individuals in the control and experimental groups, respectively, and $\mu$ represents the mean outcome across all individuals (in the population from which both groups were sampled). The error terms, $e_i^C$ and $e_i^E$, reflect all variation across individuals that is related to the outcome other than the treatment effect. It includes both the effects of personal characteristics, which we denote by $p_i$, and truly random features, including measurement error, which we denote by $r_i$, so that for both control and experimental groups,

$$e_i^{[C,E]} = p_i^{[C,E]} + r_i^{[C,E]}, \quad (2)$$

where $p_i$ and $r_i$ are independent by definition and have means of zero.

The treatment effect, $\alpha$, is conceived as a constant increment to each individual's outcome score and is the expected value of the mean difference between the treatment and control groups, that is,

$$E[Y_i^E] - E[Y_i^C] = \alpha.$$

Since the same effect, $\alpha$, is added to the outcome of each individual in the experimental group, the mean difference between the groups, $\bar{Y}^E - \bar{Y}^C$, constitutes the relevant estimate of the treatment effect. We also note that under random assignment,

$$\text{Var}[p_i^E] = \text{Var}[p_i^C] = \sigma_p^2$$

where $\sigma_p^2$ is the outcome variance attributable to person-specific characteristics. Similarly, assuming no interaction of treatment with the measurement process, it also follows that

$$\text{Var}[r_i^E] = \text{Var}[r_i^C] = \sigma_r^2$$

where $\sigma_r^2$ is true random variation. It follows then that

$$\text{Var}[Y_i^E] = \text{Var}[Y_i^C] = \sigma_p^2 + \sigma_r^2.$$

Although ideal for purposes of statistical analysis, this model for treatment effects (Equation 1) is not very realistic in many situations. Why should every individual receive an identical boost? Surely some individuals must gain more than others from experiencing the treatment.

This objection suggests that we generalize the model to represent individualized treatment effects. The model for the control group remains as before:

$$Y_i^C = \mu + p_i^C + r_i^C. \quad (3)$$

In the treatment group, we now have

$$Y_i^E = \mu + \alpha_i + p_i^E + r_i^E, \quad (4)$$

where $\alpha_i$ represents the effect of the treatment on individual $i$. The treatment effect is now a random variable that has a marginal distribution with some mean, $\mu_\alpha$, and dispersion, $\sigma_\alpha^2$. It is reasonable to assume that $\alpha_i$ will often depend upon the person-specific features represented by $p_i$. As a consequence, $\alpha_i$ and $p_i$ covary, that is,

$$\sigma_{\alpha p} \neq 0. \quad (5)$$

Now, the expected value of the mean difference between groups is

$$E[Y_i^E] - E[Y_i^C] = \mu_\alpha.$$

As a result, the observed mean difference between the groups provides an unbiased estimate of the mean effect, $\mu_\alpha$. As for the variances,

$$\text{Var}[Y_i^C] = \sigma_p^2 + \sigma_r^2, \quad (6a)$$

as before, but now,

$$\text{Var}[Y_i^E] = \sigma_\alpha^2 + 2\sigma_{\alpha p} + \sigma_p^2 + \sigma_r^2. \quad (6b)$$

Thus, as soon as we allow treatment effects to have a distribution, heterogeneity of variance across groups will occur. Included in this heterogeneity is the linkage between person characteristics, $p_i$, and the treatment effect, $\alpha_i$. Thus, in randomized experiments, heterogeneity of variance between groups can be viewed as an indicator that interaction effects of treatment with subject-specific characteristics are likely in the data.[1]

### *Simple Case*

In order to clarify the empirical consequences of individualized treatment effects, we consider the simplest case, in which both $p_i$ and $\alpha_i$ depend on just one measurable characteristic, $X_i$. In research on cognitive development, for example, $X_i$ might be a measure of cognitive ability. Thus, in this simple case, we model the person and treatment-specific effects as

$$p_i^{[E,C]} = \beta_1 X_i^{[E,C]}, \quad (7a)$$

and

$$\alpha_i = \mu_\alpha + \beta_2 X_i^E, \quad (7b)$$

where without loss of generality we will assume that $X_i^{[E,C]}$ have means of zero.

The $\beta_1$ parameter in Equation 7a captures the structural relationship between aptitude and achievement that exists in the absence of the experimental intervention. The allocation of treatment effects across individuals is represented through $\beta_2 X_i^E$. When $\beta_2$ is positive, the larger treatment effects are allocated to the higher ability students. Conversely, when $\beta_2$ is negative, the lower ability students receive the bigger effects. Substituting Equation 7 into Equations 3 and 4, we now have

$$Y_i^C = \mu + \beta_1 X_i^C + r_i^C \quad (8a)$$

and

$$Y_i^E = \mu + \mu_\alpha + (\beta_1 + \beta_2) X_i^E + r_i^E. \quad (8b)$$

Equation 8 demonstrates that the individualized treatment effects, modeled in Equation 7, imply a Treatment × Ability interaction effect.

[1] We note that heterogeneity of variance can also result from differential measurement error across groups. The presence of floor or ceiling effects on a test, for example, could cause this to occur. Even in this case, however, identifying heterogeneity of variance is important, because it indicates a model misspecification. Failure to take this into account would result in a biased estimate of treatment effects.

If the data were analyzed, however, in accordance with the conventional analysis of variance (ANOVA) model (Equation 1), we would find that

$$\text{Var}[Y_i^C] = \beta_1^2 \text{Var}(X_i^C) + \sigma_r^2 \tag{9a}$$

and

$$\text{Var}[Y_i^E] = (\beta_1 + \beta_2)^2 \text{Var}(X_i^E) + \sigma_r^2. \tag{9b}$$

Because we are assuming experimental conditions in which individuals are randomly assigned to groups,

$$\text{Var}(X_i^C) = \text{Var}(X_i^E) = \text{Var}(X). \tag{10}$$

Nevertheless, heterogeneity of variance still results, because

$$\text{Var}[Y_i^E] - \text{Var}[Y_i^C] = \beta_2(2\beta_1 + \beta_2)\text{Var}(X). \tag{11}$$

We assume, without loss of generality, that $X$ is scaled so that $\beta_1$ is positive. Then when $\beta_2$ is positive, the variance in the treatment group is larger than that in the control group. We call this a *disequalizing* situation, in that the bigger effects are allocated to the higher-ability students. Conversely, when $-2\beta_1 < \beta_2 < 0$, the variance in the treatment group is smaller than that in the control group. We call this the *equalizing* case, in which the bigger treatment effects are allocated to the lower-ability students. Finally, there is an anomalous case, when $\beta_2 < -2\beta_1$, in which again the treatment effect is disequalizing. Although mathematically possible, this condition seems unlikely to arise. Thus, a larger variance in the treatment group will most often indicate a disequalizing allocation of individual treatment effects. A smaller variance in the treatment group always indicates an equalizing allocation.

Suppose an analysis of covariance (ANCOVA) were performed instead of the standard ANOVA for the one-factor experimental design. If the sample sizes in the treatment and control groups are equal, then

$$\beta_{\text{ANCOVA}} = (\beta_C + \beta_E)/2 = (2\beta_1 + \beta_2)/2. \tag{12}$$

As a result, the true regression coefficients within each group deviate from $\beta_{\text{ANCOVA}}$ by the same amount:

$$|\beta_C - \beta_{\text{ANCOVA}}| = |\beta_E - \beta_{\text{ANCOVA}}| = |(\beta_2/2)|. \tag{13}$$

It follows that the residual variances computed on the basis of the ANCOVA model would be identical. Specifically,

$$\text{Var}[Y_i^E] = \text{Var}[Y_i^C] = (\beta_2/2)^2 \text{Var}(X) + \sigma_r^2. \tag{14}$$

Here is a case in which the model is misspecified (heterogeneity of regression is ignored), and yet the heterogeneity of variance has disappeared. This result actually constitutes a special case that occurs when there are two groups of equal sample size. If the sample sizes vary or if there are more than two groups, heterogeneity of variance will generally accompany heterogeneity of regression (assuming that the varying slopes are not specified in the model).

### *General Case*

We now extend the modeling of person- and treatment-specific effects to the multivariate case. (Hereinafter, we suppress the E and C sub- and superscripts in the interest of simplicity.) Let

$$p_i = \sum_{j=1}^{J} \beta_{1j} X_{ij} \quad \text{for} \quad j = 1, \cdots, J \text{ variables}, \tag{15}$$

and

$$\alpha_i = \mu_\alpha + \sum_{j=1}^{J} \beta_{2j} X_{ij}. \tag{16}$$

Equation 15 models the underlying structural processes that form each individual's status in the absence of a treatment intervention. Equation 16 indicates how individual treatment effects are allocated with regard to the person-specific factors captured in the $X_{ij}$ variables. This model is really quite general in that any $\beta$ coefficient or subset of coefficients may be set equal to zero. If a $\beta_{2j}$ coefficient is zero, this means that the treatment effects are being distributed without regard to that factor. If an element in $\beta_{1j}$ is zero, but $\beta_{2j}$ is nonzero, this means that the treatment has introduced a factor into the allocation process upon which natural development does not depend. For example, suppose we are comparing the effectiveness of an aural method of foreign language instruction with that of a more traditional approach. Although auditory acuity may play a negligible role in traditional instruction, it could be a very important factor for aural instruction. As a result, treatment effects are allocated as a function of what was previously an extraneous factor.

Without loss of generality, we assume that all of the $X_{ij}$ s are scaled such that in the control group the correlation between each $X_{ij}$ and $Y$ is positive. As a result, positive elements in $\beta_{2j}$s indicate that the treatment is disequalizing with regard to those factors, that is, that the treatment is amplifying preexisting differences among individuals. Conversely, negative elements in $\beta_{2j}$ s indicate equalizing effects (again discounting reverse distribution as unlikely).

Assuming that we analyze these data in accordance with Equation 1, we would find that

$$\begin{aligned}\text{Var}[Y_i^E] - \text{Var}[Y_i^C] &= \sigma_\alpha^2 + 2\sigma_{p\alpha} = \sum_{j=1}^{J} \beta_{2j}^2 \text{Var}(X_j) + 2\left[\sum_{j=1}^{J} \beta_{1j}\beta_{2j}\text{Var}(X_j)\right] \\ &\quad + \sum_{j}^{J}\sum_{j'}^{J} (\beta_{1j}\beta_{2j'} + \beta_{2j}\beta_{1j'} + \beta_{2j}\beta_{2j'})\text{Cov}(X_{j'}, X_j)\end{aligned} \tag{17}$$

for all $j \neq j'$. If all of the $\beta_{2j}$ s are positive, Equation 17 must also be positive. In the presence of such disequalizing effects, the outcome variance will be greater in the treatment than in the control group. Conversely, in the presence of pure equalizing effects, that is,

$$-2\beta_{1j} < \beta_{2j} < 0 \quad \text{for all } j \tag{18a}$$

with

$$\text{Cov}(X_j X_{j'}) \geq 0 \quad \text{for all} \quad j \neq j', \tag{18b}$$

the outcome variance will be smaller in the treatment group than in the control group. Clearly, there are also many cases between these two extremes with some $\beta_{2j}$ s positive and others negative. The overall net effect can be deduced from a comparison of group variances. When a treatment group's variance is smaller, this means that the net result of the process that allo-

Table 1
*Basic Statistics and Key Results From Gagné and Gropper Study*

| | Retention | | | | Correlation of retention | | | |
|---|---|---|---|---|---|---|---|---|
| Group | $n$ | $m$ | $s^2$ | $d$ | Ability | V/S | Pre-rate | Pre-achievement |
| Verbal | 42 | 33.5 | 96.04 | 4.570 | .22 | .28 | .02 | .09 |
| Visual | 46 | 37.8 | 75.69 | 4.348 | .16 | −.12 | .19 | .36 |
| Control | 45 | 32.2 | 187.69 | 5.156 | .33 | .07 | .27 | .23 |

*Note.* V/S = verbal–spatial ability. Heterogeneity of variance test (retention): $(1/2)\sum_{j=1}^{3} v_j (d_j - \bar{d})^2 = 9.807$, $\chi^2(2)$, $p < .01$. This test is based on the log transformation of the standard deviation, $d_j = \ln(s_j^2) + 1/v_j$. (See Equations 19 through 21.)

cates treatment effects is to reduce existing differences among individuals. In contrast, when the treatment group's variance is elevated, the allocation process is amplifying these differences. Included in the latter case is the possibility that the treatment is activating new factors previously unrelated to the outcome of interest.

## Two Illustrations from the Literature

### *Aptitude × Treatment Study of Verbal and Visual Methods of Instruction*

Our first example is an Aptitude × Treatment interaction (ATI) study (Gagné & Gropper, 1965) that was subsequently reanalyzed by Cronbach and Snow (1977). The primary purpose of the investigation was to test the hypothesis that the addition of visual illustrations to text would reduce the effects of general ability on the learning of verbal lessons. The central instruction was the same for all subjects, consisting of seven self-paced programs/lessons on mechanical advantage. Subjects were randomly assigned to three groups: visual, verbal, and control. The two experimental groups were given special introductions to each of the seven lessons. The visual group saw film demonstrations of basic concepts. For the verbal group, the same demonstrations were described only in words. The final outcome variable was achievement retention (retention) on a test 1 month after the end of the lesson. Interim outcomes included time to work through the lesson (rate) and the achievement score immediately following the lesson (achievement). General ability (ability) and verbal–spatial ability (V/S) were assessed prior to the commencement of instruction. Each group undertook two preliminary programmed lessons that provided necessary background information on mechanical advantage. This phase was identical for all three groups. These lessons provided additional aptitude measures of prerate and preachievement.

Gagné and Gropper (1965), using simple correlations and blocked ANOVA, found no significant ATI effects. Their report ended on a rather pessimistic note. They commented that there was "no reason why ATI effects could not have been revealed [in this study], if they truly existed" (p. 19) and further concluded that the ATI approach to studying learning was not a particularly promising one. The Cronbach and Snow (1977) reanalysis, employing more powerful regression techniques, detected some evidence of ATI effects on retention, although they cautioned about the suggestive character of those results. Table 1 displays key statistics reported in the Cronbach and Snow reanalysis of the Gagné and Gropper data.

Although Cronbach and Snow (1977) noted the heterogeneity of variance among the experimental groups, neither pair of investigators recognized the substantive implications of this empirical result. As was demonstrated in the previous section, the presence of heterogeneity is indicative of the fact that significant interaction effects occurred in this experiment. Further, we see from Table 2 that both treatments are variance-reducing in comparison with self-paced instruction. This indicates that the effects of self-paced instruction were disequalizing in relation to the other two methods considered in this study. Pure self-paced learning amplified differences among students that the supplemental direct instruction attenuated.

Cronbach and Snow's (1977) reanalysis fitted separate models for regressing retention on ability, V/S, achievement, and preachievement for each group. (The residual variances from this conditional model are displayed in Table 2). After controlling for these four variables, the variance difference between the verbal treatment and the self-paced instruction was no longer significant. The residual variance in the visual group, however, remains significantly smaller. This implies that there are still unspecified variables, the effects of which operate differently under visual instruction than under other methods. Whatever these specific variables are, it is clear that visual methods reduce the effect of unmeasured individual differences, differences that are amplified under both the verbal and the self-paced forms of presentation.

Thus, by a careful consideration of variance differences across the experimental groups, we come to a considerably different conclusion than that reached by the original authors. Quite contrary to Gagné and Gropper's (1965) rather pessimistic ending, the evidence assembled in their study suggests that the visual form of instruction is a very promising method. The level of retention was higher and the effects distributed in a relatively equalizing fashion.

### *Teacher Expectancy Experiment*

Our second illustration draws on results from an experiment conducted by Kester (1969), and later published by Kester and

Table 2
*Residual Variances in Retention (After Controlling for Ability, V/S, Prerate, and Preachievement)*

| Group | $s^2_{y\|x}$ | $d$ |
|---|---|---|
| Verbal | 65.74 | 4.030 |
| Visual | 23.47 | 3.180 |
| Control | 78.83 | 4.392 |

*Note.* V/S = verbal–spatial ability. One degree of freedom contrasts: $H_0: \ln \sigma_{\text{verbal}}^2 = \ln \sigma_{\text{control}}^2$; $z = (4.030 - 4.392)/[2(1/37 + 1/40)]^{1/2} = -1.12$, *n.s.*
$H_0: \ln \sigma_{\text{visual}}^2 = (1/2)(\ln \sigma_{\text{verbal}}^2 + \ln \sigma_{\text{control}}^2)$; $z = (3.180 - 4.211)/[2/41 + 1/4(2/37 + 2/40)]^{1/2} = -2.47$, $p < .01$. These tests are based on the procedure for examining linear contrasts given in Equation 22.

Letchworth (1972). The study assessed the effects of experimentally induced teacher expectancies on pupil IQ. Within each of 24 classrooms, several students were assigned at random to either a high-expectancy or a control condition. In all, 75 students were assigned to each group, though data for one control student was lost. The authors found no effect of teacher expectancy on pupil IQ.

In this study, too, the treatment had an effect on the variance that went unrecognized in the original investigation. As is indicated in Tables 3 and 4, a reanalysis of the Kester data using preexperiment IQ as a covariate reveals substantial differences in the residual variance in the experimental and control groups.[2] Unlike our first example, the treatment apparently exerts a disequalizing effect here. Specifically, the residual variance in the experimental group (56.52) was significantly larger than the residual variance in the control group (32.59), $F(73, 74) = 1.73, p < .02$.

Two different explanations are possible for this result. First, it might be that teachers differed in their response to the expectancy-inducing information. If some teachers acted on the basis of the inflated expectancies while others simply ignored them, a larger treatment group variance would result. Alternatively, the effect of the treatment might depend on student characteristics (e.g., varying individual responsiveness or needs for praise and reinforcement). This, too, could produce heterogeneity of variance.

Further examination of the data sheds considerable light on the tenability of these alternative explanations. If, indeed, teachers responded differentially to the expectancy-inducing information, we would expect to find evidence that the magnitude of the treatment effect varied across classrooms. This hypothesis can be examined by introducing classrooms as an additional factor in the ANCOVA, with attention focusing on the Treatment × Classroom interaction effect. The 2 × 24 (Treatment × Classroom) ANCOVA, however, revealed no evidence of a significant interaction, $F(1, 23) = .83$.[3] This result indicates that the source of the heterogeneity is within classrooms, a result consistent with our second explanation. Specifically, the second explanation, that treatment effects depend on student characteristics, implies that the within-classroom variance would be larger for the treatment group than for the control group. In fact, this is exactly what occurs (see Table 3).[4] The within-classroom variance was 54.69 in the experimental group and only 25.24 in the control group, yielding an experiment-to-control variance ratio of 2.17, $F(51, 50) = 2.17, p < .01$. Hence, the effect of the treatment did indeed depend on unmeasured student characteristics.[5]

Table 3
*Basic Statistics and Key Results From Kester (1969) Study*

| Statistic | Experimental | Control |
|---|---|---|
| Pretest IQ $m_x$ | 101.11 | 100.07 |
| Posttest IQ $m_y$ | 104.62 | 102.47 |
| Pretest $s_x^2$ | 30.36 | 28.20 |
| Posttest $s_y^2$ | 71.40 | 47.47 |
| Residual variance, $s^2_{y\|x}$ | 56.52 | 32.59 |
| Pooled within-classes residual variance, $s^2_{y\|x,\text{classes}}$ | 54.69 | 25.24 |
| Sample sizes, $n$ | 74 | 75 |

Table 4
*Treatment × Classroom ANCOVA (With Pretest IQ as a Covariate)*

| Source | *df* | *SS* | *MS* | *F* | *p* |
|---|---|---|---|---|---|
| Covariate | 1 | 1,147.91 | 1,147.91 | | |
| Treatment | 1 | 83.28 | 83.28 | 2.54 | .07 |
| Teachers | 22 | 1,863.24 | 84.69 | | |
| Treatment × Teachers | 22 | 721.91 | 32.81 | .83 | n.s. |
| Residual | 102 | 4,021.97 | 39.43 | | |

Here, too, our framework for studying variance heterogeneity enabled us to take a study originally dismissed as producing null findings and to discover a potentially important result. The effects of teacher expectancies, rather than being negligible, were variable. Moreover, we were able to dismiss teacher differences as the source of this variability and to conclude instead that the treatment effects depended on unidentified student characteristics. Thus, the results of this simple experiment suggest that a more sophisticated subsequent study be undertaken to identify the precise student characteristics involved and thereby to contribute to a better understanding of the mechanism by which teachers' expectations differentially affect their students.

## Techniques for Modeling Variances

### *Normal Theory Methods*

Clearly, a careful examination of variance should be a routine component in the analysis of data from psychological experi-

[2] This first analysis was based on a simple treatment–control group design (ignoring classrooms), using pretreatment IQ as a covariate. Specifically, the model was $Y_{ij} = \mu + \alpha_j + \beta(X_{ij} - \bar{X}_{ij}) + e_{ij}$ where $\mu$ is the grand mean, $\alpha_j$ is the group effect (treatment versus control), $\beta$ is the regression coefficient pooled within the treatment and control groups, and $X_{ij}$ is the pretreatment IQ. The use of the pooled within-group coefficient was justified in that the separate regression coefficients in the treatment and control groups were not statistically different. The parameter estimates from this ANCOVA model were used to generate predicted outcomes, $\hat{Y}_{ij}$, for each subject, and a residual variance, $\sum_{j=1}^{n_j} (Y_{ij} - \hat{Y}_{ij})^2/(n_j - 1)$, computed separately for the treatment and control groups.

[3] The model for this second analysis was $Y_{ijk} = \mu + \alpha_j + \delta_k + \gamma_{jk} + \beta(X_{ijk} - \bar{X}_{ij}) + e_{ijk}$ where $\alpha_j$ denotes the treatment effect (experimental vs. control); $\delta_k$ is the effect of classroom $k$, $(k = 1, \cdots, 24)$; $\gamma_{jk}$ represents the Treatment × Classroom interaction effect; and $\beta$ is the regression coefficient pooled within the 48 Treatment × Classroom cells.

[4] As in the first analysis, parameter estimates based on this model were used to generate predicted outcomes $\hat{Y}_{ijk}$, and a residual variance was computed separately for experimental and control groups using the formula $\sum_j \sum_k (Y_{ijk} - \hat{Y}_{ijk})^2/(\sum_j \sum_k n_{jk} - 24)$, where $n_{jk}$ is the number of children in treatment $j$, classroom $k$.

[5] This analysis treated both classrooms and treatments as fixed factors. As a reviewer pointed out, the optimal analysis of these data is

ments. To facilitate future efforts of this sort, we review in this section basic statistical techniques for analyzing variances and testing hypotheses about them. Although this theory is well established in the statistical literature (see, e.g., Miller, 1986), it is not widely known by practicing researchers.

The method for comparing the variances of two independent groups is well-known. Assuming the data are normally distributed, the ratio of two sample variances is distributed as an $F$ statistic with $v_1$ and $v_2$ $df$s respectively, under the homogeneity of variance hypothesis. This is the standard parametric test for comparing variances in two groups. Extension to more than two groups is not direct, however, because there is no simple statistical theory for linear modeling of sample variances. The most common alternative is to transform the sample variances, $s_j^2$, for each of the $J$ groups such that the transformed statistics are approximately normally distributed. Estimation- and hypothesis-testing techniques from normal distribution theory are then applied.

Several normalizing transformations of $s_j^2$ have been suggested in the literature, including the log transform and the square root and cubed root of $s_j^2$ (see Kendall & Stuart, 1969; or for a more recent review, Raudenbush & Bryk, 1987). Although the cube root transform converges very quickly to normality and thus offers distinct advantages with very small sample sizes, the log transformation is generally preferable for several reasons. First, linear contrasts among the ln $(s_j^2)$ are invariant to changes of scale in the raw data (Box & Tiao, 1973). For instance, standardizing the data around the grand mean for several groups has no effect on estimated contrasts among the log-transformed variances. This invariance property does not hold $s^2$, $s$, or the cube root transform. Second, when the raw data are normal, the ln $(s_j^2)$ are approximately normally distributed and with stable variance. The approximate sampling variance for ln $(s_j^2)$ is $2/v_j$, which does not depend on the population variance, $\sigma^2$. Third, a bias correction factor for ln $(s_j^2)$ can be introduced that improves the accuracy of the asymptotic approximation. Raudenbush and Bryk (1987) demonstrate that this bias-adjusted log transform is an excellent approximation with sample sizes as small as 10 per group, which covers most of the cases likely to be encountered in social and behavioral research.

Specifically, we define the transformation

$$d_j = \ln (s_j^2) + 1/v_j, \tag{19}$$

where $v_j$ is the degrees of freedom in group $j$ and $1/v_j$ is the bias correction factor. These transformed variances can then be used in any standard linear model technique.

For example, as first suggested by Bartlett and Kendall (1946), we can estimate the residual variance separately for each cell in an ANOVA design and perform an ANOVA on the transformed variance statistics. A simple application is the omnibus test for heterogeneity of variance, as was previously used in Table 1. For an ANOVA design, if we define

$$\bar{d} = \sum_{j=1}^{J} v_j d_j / \sum v_j, \tag{20}$$

where the summation is taken across all of the cells of the design, $j = 1, \cdots, J$, then the test statistic for the omnibus homogeneity of variance hypothesis is

$$(1/2) \sum_{j=1}^{J} v_j (d_j - \bar{d})^2, \tag{21}$$

which has an approximate chi-squared distribution with $J - 1$ degrees of freedom.

Because the $d_j$ are approximately normally distributed, we can also perform both simple and complex contrasts among specific cell values. In general, for any linear contrast,

$$\sum_{j=1}^{J} c_j d_j \quad \text{with} \quad \sum c_j = 0$$

among the $J$-transformed variances, the test statistic

$$\sum_{j=1}^{J} c_j d_j / (2c_j^2/v_j)^{1/2} \tag{22}$$

follows a $z$ distribution under the null hypothesis. This test was also used in Table 1, in which we separately compared the verbal and visual methods of instruction with the control condition. In general, any standard multiple comparison procedure can be applied to the $d_j$ statistics. For a further discussion and illustration of these techniques, see Games (1978a, 1978b). His development is identical to ours except that he does not take into account the bias correction factor in Equation 19.

In fact, modeling of variances can be approached as a general linear model problem. Specifically, Raudenbush and Bryk (1987) have formulated a mixed model for sample variances utilizing the log-transformed $d_j$ introduced above. These $d_j$ are represented as a linear function of a set of fixed effects including possible design variables and of random effects that might result from sampling units such as schools, classrooms, or persons.

### *Sensitivity to Distributional Assumptions*

Parametric test statistics for variances are not robust to violation of the normality assumption for the raw data. In particular, the distribution of the test statistics are sensitive to kurtosis in the parent distribution. If the raw data have "fat tails" (a platykurtic distribution), the true $\alpha$ levels will be underestimated and the probability of a Type I error increased. With leptokurtic data, the reverse is true (for a review, see O'Brien, 1979).

Table 5 displays the kurtosis values, $\gamma_2$, of several common distributions. For each distribution, we present a correction factor that can be used in two different ways to guard against invalid inferences.

First, if the underlying distribution is known or can be approximated, standard errors can be adjusted for kurtosis. Suppose, for example, that the outcome variable is a behavioral count, such as the frequency of student-initiated interaction during a class, and that students initiate an average of one interaction each. This outcome should be well represented as a Pois-

based on a mixed model with treatments fixed and classrooms random. However, the $F$ test for the Treatment × Classroom interaction is the same using both models if the design is balanced (Kirk, 1982, p. 391). Given that the Kester data were nearly balanced and that the $F$ test was trivially small, no further analysis was needed.

Table 5
*Correction Factors for Several Distributions*

| Distribution | $\gamma_2$ = kurtosis | Correction factor $C = (1 + \gamma_2/2)^{1/2}$ |
|---|---|---|
| $t$, $df = 5$ | 6.00 | 2.00 |
| $t$, $df = 6$ | 3.00 | 1.58 |
| $t$, $df = 10$ | 1.00 | 1.22 |
| $t$, $df = 20$ | .37 | 1.09 |
| $t$, $df = 30$ | .23 | 1.06 |
| $t$, $df = 60$ | .11 | 1.03 |
| Poisson, $\mu = 1$ | 1.00 | 1.22 |
| Poisson, $\mu = 2$ | .50 | 1.12 |
| Poisson, $\mu = 3$ | .33 | 1.08 |
| Normal | 0.0 | 1.00 |
| Unimodal, symmetric likert | .50 | .87 |
| Beta ($p = 2$, $q = 2$) | −.86 | .76 |
| Uniform (or uniform likert) | −1.30 | .59 |

*Note.* For the $t$ distribution, $\gamma_2 = 6/(df - 4)$. For the Poisson distribution, $\gamma_2 = \mu - 1$. The Poisson is a sensible model for the probability of $n$ events in a unit interval of time, where $E(n) = \mu$. Examples might include student-initiated interactions or days absent. The unimodal symmetric Likert has five categories with probabilities of .1, .2, .4, .2, and .1, respectively. The beta distribution with $p = 2$, $q = 2$, is unimodal, symmetric, and truncated so that the variable takes on values between 0 and 1. The kurtosis depends on the fourth moment of the data distribution and indicates the density of observations in the tails of the distribution. See Johnson and Kotz (1970).

son variate with mean, $\mu = 1$. For this case, the correction factor is $C = 1.225$. The value of the $z$ statistic is simply divided by $C$ to obtain a test statistic that is corrected for the kurtosis in the parent distribution. Without the correction, the $z$ test would be too liberal, resulting in elevated Type I errors. Note that if there were an average of three interactions per student (i.e., $\mu = 3$), then the correction factor, $C$, would be 1.080. Alternatively, suppose the outcome is a Likert scale with an approximately uniform distribution of responses. Now the appropriate correction factor would be .59. Without this correction, the $z$ test would be too conservative.

The second way of using these correction factors is as a form of sensitivity analysis. One simply calculates the value of the correction factor needed to reverse an inference. The question becomes, Is it possible that the data were actually generated from such a distribution?

For example, consider the analysis presented in Table 3 for the Kester study. Assuming normality, the $z$ test of the difference between experimental- and control-group variances (pooled within classrooms), is

$$z = \frac{\log(54.69) + 1/50 - \log(25.24) - 1/51}{\sqrt{2/50 + 2/51}} = 2.75,$$

$p < .005$. Because the critical value of $z$ at $\alpha = .05$ is 1.96, the correction factor needed to overturn the inference is 2.75/1.96 = 1.40. As Table 5 shows, this correction factor is associated with a rather fat-tailed distribution, a $t$ distribution with approximately 8 degrees of freedom. However, examination of a normal probability plot actually showed a distribution with "thin tails" (so that the correction factor is likely to be smaller than unity). Thus, the sensitivity analysis supports the normal-based inference.

When nonnormal data occur, nonparametric tests for variances represent another option (for a review, see Miller, 1986, chap. 7). One fairly flexible technique, discussed by O'Brien (1979), is a generalization of the Scheffe test (see Glass & Hopkins, 1984, p. 356) that involves use of jackknife-type estimators. This approach is more complex computationally, and as O'Brien notes, the variance of these estimators depends upon their means, and this dependence can be problematic.

## Discussion

Both of the examples presented in this article were chosen to illustrate an important point. Substantively significant empirical results have been ignored because research methodology has tended to focus exclusively on mean differences. Standard methodological training has left researchers largely unaware of the theoretical significance of variance heterogeneity, partly because basic texts tend to view such heterogeneity as a methodological nuisance rather than a source of important information.

The practice of routinely searching for data transformations that will eliminate heterogeneity is misguided. Although such transformations may be warranted, their legitimacy derives from purely substantive considerations; that is, a transformation is justified only if the transformed metric is more meaningful than the original metric. Variance, stabilizing transformations are necessarily nonlinear transformations. Hence, the original and transformed metrics cannot both be interval measures of the same construct. Because linear model analyses require the outcome to be measured on an interval scale (at least approximately so), a variance-stabilizing transformation is justified only if the transformed metric approximates an interval scale measure of the construct better than does the original metric.

Most analyses of experimental data assume that the treatment is a fixed entity that can be formally defined and uniformly administered to each individual. In such a case, the only source of heterogeneity of variance is individual differences in responsiveness to the fixed treatment. Such differences constitute interaction effects between person characteristics and treatment group membership.

In many research contexts, however, the treatment that is actually implemented may vary across individuals. We hypothesized such effects, for example, in considering the heterogeneity in Kester's (1969) teacher expectancy study. More generally, research on instruction often utilizes several classrooms, therapy groups, or other groupings in order to obtain a sufficient subject sample. It is reasonable to assume, however, that the different teachers in these classrooms will vary in their use of the instructional intervention. The resultant variations in the treatment implementation as well as differences in individual subjects characteristics can produce variance in the treatment effects.

One obvious response to our objections to the constant treatment effect model (Equation 1) is that although this model is not literally correct, it is still useful in that it provides an adequate summary measure of a treatment's overall effect. Knowledge that a treatment works well on the average is considered to

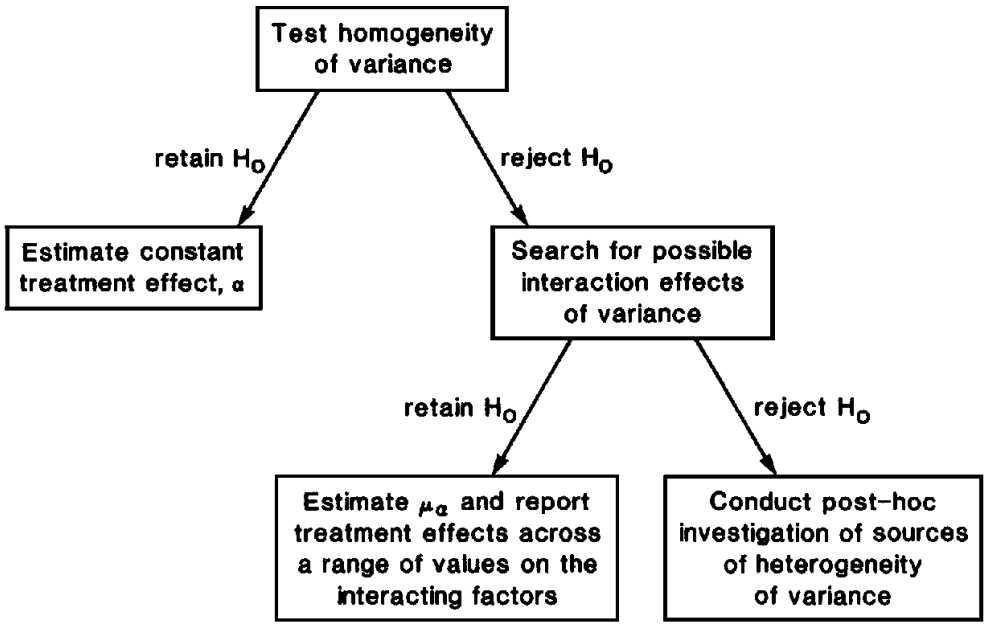


*Figure 1.* Decision tree in analyzing experimental data. (It is good data analysis practice to consider the tenability of the normality assumption with conventional techniques such as probability plots. If the data appear nonnormal, the analyst should examine the sensitivity of the homogeneity tests to the likely violations of the normality assumption. See text.)

be useful. Embedded in this interpretation of mean differences between groups is an implicit value judgment about the relative worth of differential effects. For example, we are assuming that five individuals gaining 2 points is an equal counterweight to one individual losing 10 points. If, however, the 10-point loss reflects catastrophic consequences for that individual, the value of these two sets of consequences may be far from counterbalancing. For clarification, the concerns we are raising now are not measurement issues but rather questions of value. These problems would exist even if our measures had ideal psychometric properties, such as a perfect interval scale.

On a more profound level, we assert that psychological research will be better served by assuming a priori that treatment effects are random rather than fixed. Here the treatment effects are random not because the treatments themselves constitute a random sample from a population of treatments but because the number of subject characteristics that may interact with treatments is so vast that one cannot assume a priori that our statistical models can specify all relevant interactions. If the treatment effect is therefore a truly random variable, the goal of statistical inference is to learn about its distribution. Technically, the distribution of interest is the conditional distribution of the treatment effects given the characteristics of individuals, settings, and characteristics of treatment implementations. Important parameters of this distribution include conditional means (specified by main effects and interactions) and conditional variances.

In summary, the presence of heterogeneity of variance across treatment groups is a strong indicator that the treatments have differential effects on individuals. In the presence of such individualized effects the mean treatment effect, $\mu_\alpha$, provides an insufficient summary. The sources of this heterogeneity should be identified and treatment effects reestimated conditionally on the identified interactions. Specifically, we propose a sequence of analytic activities, as is displayed in Figure 1.

Testing the homogeneity of variance hypothesis should be a routine component in the analysis of experimental data. A finding of homogeneity of variance across groups indicates that the classical treatment effect model is tenable. The analyst may safely proceed to estimate $\alpha$ as the appropriate statistical summary of the treatment effect. If the homogeneity hypothesis is rejected, the analyst should consider the possibility of interaction effects in the experiment. Assuming that such factors are identified, the analyst should fit a new model that incorporates the interaction terms and test the homogeneity of residual variances. If the latter is sustained, the investigator should report both the average effect, $\mu_\alpha$, and estimated effects across a range of values on the interacting factors. A comprehensive description of the multivariate distribution of treatment effects becomes the central scientific goal.

Failure to sustain a homogeneity of variance hypothesis means that there are still unidentified sources of individual variability present in the study that may confound the interpretation of estimated treatment effects. This is an important finding that requires explicit acknowledgment and further investigation. Post hoc studies, as we have illustrated in our two examples, may help to reveal the source of the unmeasured interaction effects (e.g., interactions with subject characteristics) and their general nature (e.g., equalizing versus disequalizing). Such exploratory analyses can suggest possible explanations for the observed heterogeneity that should be formally tested in subsequent research.

More generally, the possibility of individualized treatment effects in experimental studies has implications for research design and reporting of results. Data on personal characteristics that might interact with treatment effects should be collected. Both existing theory and careful observation of the experiment in progress may help to direct this extended inquiry. Should heterogeneity of variance be encountered, these additional data can be explored for possible explanations using the methods presented in this article.

The presence of individualized treatment effects also has important implications for subsequent meta-analysis of research programs. Because studies rarely employ a random sampling of both person and treatment characteristics, differences across studies in these factors will produce heterogeneity of mean effect sizes. At a minimum, studies should report a full elaboration of subject and treatment characteristics in order to facilitate subsequent synthesis. Ideally, the focus of the meta-analysis task should also be redirected. Attention should shift from synthesizing a single effect estimate, that is, $\mu_\alpha$, from each study to summarizing the distribution of the treatment effects conditionally on subjects, settings, and treatment characteristics. Both conditional means and conditional variances are relevant to the summary. Given that each study would produce a vector of parameter estimates, a multivariate approach to meta-analysis is needed. Statistical methods for such multivariate meta-analyses have begun to appear (Hedges & Olkin, 1985; Raudenbush, Becker, & Kalaian, 1988; Rosenthal & Rubin, 1986), and further work in this area is warranted.

Finally, although the research designs considered here are simple, the methods readily extend to more complex cases, including multifactor models having within- and between-subject components. Raudenbush (1988) provided methods for estimating and testing contrasts among correlated dispersion estimates. This theory enables study of the effect of within-subjects treatments on variances. The sequence of steps presented in this article apply again: If variances are found to be heterogeneous, model variation among the variances first; then estimate mean effects and report differential effects of treatments for subjects of differing background.

## References


Bartlett, M. S., & Kendall, D. G. (1946). The statistical analysis of variance-heterogeneity and the logarithmic transformation. *Journal of the Royal Statistical Society* (Suppl. 8), 128–138.

Bloom, B. S. (1984). The 2-sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4–16.

Box, G. E., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis.* Reading, MA: Addison-Wesley.

Bryk, A. S. (1978). Evaluating program impact: A time to cast away stones, a time to gather stones together. *New Directions for Program Evaluation, 1,* 31–58.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitude and instructional methods.* New York: Irvington.

Gagné, R. M., & Gropper, G. L. (1965). *Individual differences from learning in verbal and visual presentations* (ERIC No. ED010377). Unpublished report, American Institutes for Research, Pittsburgh, PA.

Games, P. A. (1978a). A three-factor model encompassing many possible statistical tests on independent groups. *Psychological Bulletin, 85,* 168–182.

Games, P. A. (1978b). A four-factor structure for parametric tests on independent groups. *Psychological Bulletin, 85,* 661–672.

Glass, G. V., & Hopkins, K. D. (1984). *Statistical methods in education and psychology.* Englewood Cliffs, NJ: Prentice Hall.

Hays, W. L. (1981). *Statistics.* New York: Holt, Rinehart & Winston.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York: Academic Press.

Johnson, N. L., & Kotz, S. (1970) *Distribution in statistics, Vols. I–II.* New York: Wiley.

Kendall, M. G., & Stuart, A. (1969). *The advanced theory of statistics* (Vol. 1). London: Hafner.

Kester, S. W. (1969). The communication of teacher expectations and their effects on the achievement and attitudes of secondary school pupils (Doctoral dissertation, University of Oklahoma, 1969). *Dissertation Abstracts International, 30,* 1434-A. (University Microfilms No. 6917653).

Kester, S. W., & Letchworth, G. A. (1972). Communication of teacher expectations and their effects on achievement and attitudes of secondary school students. *Educational Researcher, 66,* 51–55.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences.* Monterey, CA: Brooks/Cole.

Light, R. J., & Smith, P. V. (1971). Accumulating evidence: Procedures for resolving contradictions among different studies. *Harvard Educational Review, 41,* 429–471.

Miller, R. G. (1986). *Beyond ANOVA: Basics of applied statistics.* New York: Wiley.

O'Brien, R. G. (1979). A general ANOVA method for robust tests of additive models for variances. *Journal of the American Statistical Association, 74,* 877–880.

Raudenbush, S. W. (1988). Estimating change in dispersion. *Journal of Educational Statistics, 12,* 241–270.

Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin, 103,* 111–120.

Raudenbush, S. W., & Bryk, A. S. (1987). Examining correlates of diversity. *Journal of Educational Statistics, 12,* 241–269.

Rosenthal, R., & Rubin, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin, 99,* 400–406.

Winer, B. J. (1971) *Statistical principles in experimental design.* New York: McGraw-Hill.


Received April 13, 1987
Revision received February 29, 1988
Accepted March 10, 1988 ■